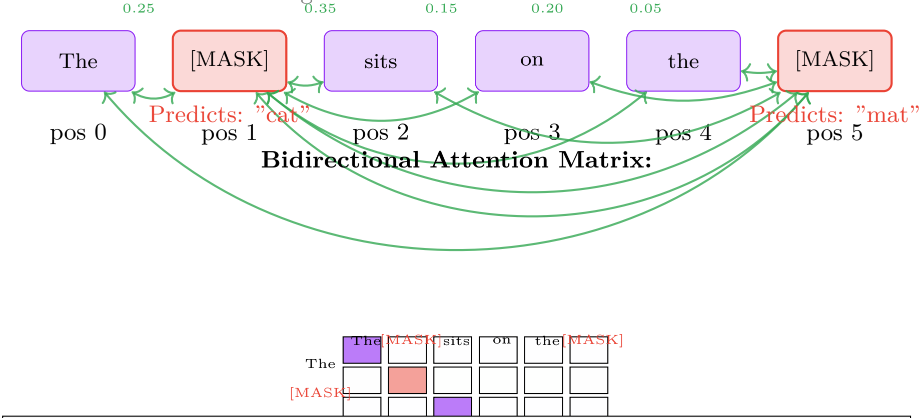
MLM Training Example:

Original: ”The cat sits on the mat”



Bidirectional Attention Matrix:

Key Differences from Autoregressive Models

Bidirectional Context: Uses both left & right context

No Causal Masking: Future tokens inform predictions

Parallel Prediction: Multiple masks predicted jointly

MLM Objective: Predict original tokens using full bidirectional context